National Cancer Informatics Program Launch (NCIP) Meeting Summary

Major Recurrent Themes

- The science underlies everything when developing computational methods and an information infrastructure. Similarly, computational biology should be driven by key biological questions. We must integrate computational biology needs with experimental biology needs. The National Centers for Biomedical Computing (NCBCs), funded as an NIH Roadmap project, exemplify the driving biological project model.

- Start small, as did the Clinical and Translational science Awards (CTSA) program, and employ a bottoms-up rather than a top-down approach to empower imagination and innovation. It is easier to achieve cultural change from the bottom up. Also, we should recognize the extraordinary difficulty of building enterprise-level applications and getting their acceptance from the top down.

- A bottom-up approach makes training and education crucial: basic scientists and clinicians need to become informatics savvy, so cancer informatics training is needed, especially for postdoctoral fellows and graduate students.

- There's a driving need for data integration at every level of the biomedical enterprise. Two related themes—the need for data sharing or exchange and interoperability—were also discussed at length. In all these areas, a robust infrastructure and standards are key. There is also a need for the integration of analysis methods. Paralleling data integration is integration of communities—informaticists, computational biologists, experimental biologists, clinical researchers—a move that requires cultural change and targeted training programs.

- The NIH Clinical and Translational Science Awards (CTSA) program was repeatedly cited as a key exemplar (*e.g.*, illustrating the principle of starting small) and as a potential partner for NCIP.

- NCIP should provide policy and infrastructure at a high level, but NOT at the implementation level (again, to allow space for bottom-up innovation). NCIP should serve a coordinating role—at the beginning, this may work best with smaller rather than larger research groups tackling specific, sharply focused research challenges.

- We must acknowledge the centrality of clinical trial participants in the clinical information infrastructure—conceive of the patient as a Cancer Information Donor. Driven in part by the ongoing evolution of meaningful use requirements for electronic medical records (EMRs) set by the Office of the National Coordinator (ONC), *i.e.*, Meaningful Use, and the concept of the learning health care system. The goal is participatory precision medicine that incorporates the patient perspective. According to information presented at a recent Institute of Medicine (IOM) meeting, 26% of the population is willing to share identified health information—thus constituting a potentially enormous resource. Much of this information will be unstructured.

- We must develop lightweight informatics solutions in response to specific scientific needs that can be developed quickly—within a matter of months—and pushed out the door; such applications must be user-friendly to facilitate rapid, widespread adoption. Lightweight development allows for early-stage "wins." Ideally, there should be a balance between lightweight and industrial-strength applications and infrastructure elements to achieve agility and adaptability.

- NCIP should partner closely with and support other NCI and NIH programs and insert informatics requirements with teeth (à la CTSA) into ongoing foundational NCI programs/grant mechanisms administered by the NCI Divisions, Offices and Centers (*e.g.*, SPOREs, R01s). NCIP is to be aligned with the mission and goals of the NCI.

Important Emerging or Secondary Themes

- Improving reproducibility with regard to *in silico* research methods is critical.

- There is a need for effective data-mining applications for both structured and unstructured text, as well as continued improvements in natural-language processing (NLP)—a key challenge, especially in clinical research. Much more unstructured clinical data is going to become available soon.

- There is a need for standards to govern data exchange; the NCI informatics infrastructure should engage more directly with ONC Health Information Exchange (HIE) standards. There were also strong recommendations that NCIP partner with the National Library of Medicine (NLM), which has established interoperability standards for its databases.

- There is a need for secure data repositories (the trust structure) and simultaneously for better methods of data reduction and other types of data management. Perhaps NCIP should establish a national data-storage center in partnership with the NLM.

Introductory session

Dr. Komatsoulis, National Cancer Informatics Program Kickoff Meeting

The objectives of this meeting are to launch the new NCI National Cancer Informatics program (NCIP) and to begin more specific discussions on current and future needs for biomedical informatics capabilities in cancer biology, clinical, and translational research, as well as needs for interoperable information systems (including vocabularies, data elements, and data standards). We need to conceptualize a governance structure for the program that will fulfill scientific needs and create a set of working groups that will study these questions in more detail. NCIP will entail four categories of support: interoperability to support data sharing (vocabularies, common data elements, data standards, technical specifications, core biomedical informatics applications, and security); biomedical informatics research and development (creation of the next generation of biomedical informatics algorithms, statistical methods and tools, etc.); *in silico* research, which uses computational rather than experimental methods; education and training (create training grants to the develop the next

generation of biomedical informatics professions and include training support to assist researchers and clinicians in becoming "informatics savvy" even if they do not become full-fledged biomedical informatics professionals).

Dr. Varmus (untitled, no slides)

Science underlies everything—both computational methods and information infrastructure. There is a need to focus on data sets: in particular, interactions between clinical and research data to establish strong relationships between basic, clinical, and translational research on the one hand and common medical practice on the other. Pursuing cancer science requires cooperation from the communities, helpful rules governing the enterprise, and empowering imaginations (ideas percolating from the bottom up). Standards are incredibly important; interoperability is essential; flexibility is required. This meeting's goals: the need for central guidance to support creative energy; using science to understand and control cancer.

National Cancer Advisory Board (NCAB) *ad hoc* informatics Oversight Committee.

Dr. Masys, Desirable Features of a National Cancer informatics Program

"Federal agencies are at their best supporting standards and recognizing bottom-up innovations."

The current status of the NCAB Informatics Working Group (IWG): review of CBIIT/caBIG contracts is about 50% complete; now undertaking concept reviews of new projects; providing a force for continuity; and will monitor and advise on the ongoing NCIP implementation and progress. The current "sense" of the IWG is that the new NCIP will need to be agile enough to respond to change; use generalizability as a key criterion for projects; recognize innovations coming from the bottom up; and support maturation/hardening of innovations. We will also have to develop a way to evaluate success or failure: what will, for example be the stopping rules for ending a project, even when it meets technical objective but fails to be adopted? Flexibility and generalizability will be key criteria. Three desiderata for the NCIP: measurable progress in contributing to cancer research or care; administrative agility (but noted standards cannot be quickly developed in a consensual, community-driven way); an objective, to-be-determined third-party mechanism for ongoing program assessment.

Session 1: Basic Science: Cancer Biology and Genomics

Presentation 1: Dr. Califano, National Cancer Informatics Program: Making Sense of Data via Models

Computational biology should be driven by key biological questions. There are enormous gaps in our knowledge; we probably know 0.1 percent—for example, there's a conspicuous gap in our knowledge of cell regulatory logic. Another pressing question in this regard is "how does genetic heterogeneity in cancer lead to a homogeneous response?". We need data- and model-driven discovery, as well as model-based meta-analysis, to provide a hypothesis-driven framework; a central problem is normalizing data and tissue samples for integrated genomics—this could employ a machine-driven approach or a Google Maps solution. Another major challenge is that data information and tissue information are not connected. Aligned to this is the need to develop approaches to unstructured data. A coordinated

paradigm for discovery would entail cultural change (integrate computational and experimental biologists). There is a need for hybrid educational programs and bringing communities together (the "humid" model). What is necessary: creating the necessary data sets (TCGA provides a model, but there's an overwhelming need for proteomics data); supporting lightweight mechanisms for integrating diverse data modalities; supporting integration of molecular and clinical data; and building the infrastructure by, for example, creating a portfolio of centers of excellence in cancer informatics. It is also important to build an infrastructure—perhaps a cloud—to handle petabytes of data; this infrastructure should be designed to reduce bottlenecks to data access. Proposal: an "Encode+ project for cancer" that would cover collection, reconstruction, interrogation, and validation. Reproducibility is critical.

Presentation 2: Dr. Sander (untitled, no slides)

There's a need to build bridges between biological data and biological knowledge—information needs to be reduced into simple biological statements. We must work harder on building data portals—similar to, for example, the cBio Portal—that provide access to multidimensional cancer data. Following on this, there is a need for data organization and efficient data transfer (interoperability and interoperation, respectively). Missing information must be addressed (for example, TCGA lacks certain types of patient information such as treatment). Natural-language processing needs to be further developed to handle unstructured clinical information. There is a strong need for synthesis rather than a Google Maps approach: for example, see the integrated genomics viewer developed at the Broad Institute.

Presentation 3: Dr. Wold, Cancer Genomics 2013-2018

We must consider what information and computational needs there will be for cancer genomics between 2013 and 2018, especially in regard to needs for bioinformatics and computational activities as well as needs for storage and user access to resources. We have to manage both structured and unstructured clinical data. Getting clinical data and metadata is a major issue: natural-language processing is crucial. Data storage and access will also prove crucial. CGHub offers only very basic storage at this time; the VA, however, has developed an impressive system for electronic medical records (EMRs). With clinical information, we must embrace the concept of the patient as Cancer Information Donor. Data abstraction/reduction (*e.g.*, lossy) will be critical for consuming clinical information. We may be reaching the point where we will keep little of the sequence information, but the EMR data will get bigger.

Presentation 4: Dr. White, NCIP Workshop

Within the context and anticipated needs arising from systems biology and the dynamics of cancer biology, there will be a scale-up from terabytes to exabytes leading to needs for data centers, elastic clouds, and high-performance computing: see Grossman and White, A vision for a biomedical cloud, *JIM* 2012, and the Bionimbus Cloud project. We need to be able to account quantitatively for the dynamic nature of biological processes such as the ways in which the tumor microenvironment changes. Sequencing data is not likely to get smaller. The volume of genome-sequencing data is growing

exponentially, and the trend looks almost certain to continue. Therefore, we need a national infrastructure to handle the volume. Requirements for a national biomedical cloud include security, in particular secure communication with private clouds; on-demand and scalable storage; on-demand and scalable analysis; scalable ingestion of data; support for data liberation; and peering with community and public clouds.

Dr. Komatsoulis, Summing-up Remarks and Ensuing Discussion

What concrete steps we can take? I propose a straw man: assess ways to integrate informatics into the core of other programs; develop a standard way of representing genetic and genomic information; achieve information reduction; provide support for a public biology cloud; create centers for integrating information over the long term.

Singer: Will NCIP support crowdsourcing and other social-networking activities? We need standards to provide a framework for people to work together—a national standard of some kind—and to establish policy. (seconded by Chin)

Unidentified Speaker: Look to the National Science Foundation and the Department of Defense for models.

Unidentified Speaker: We need a W3C-like group to develop standards for cancer biology.

Unidentified Speaker: Establishing a mechanism to maintain semantic infrastructure is important. Chartered projects would help define data integration. (seconded by Saltz)

Saltz: Regarding a public cloud, we will need to understand bulk computing and architecture needs. It is good to have driving biological problems. Research into technology is a legitimate goal.

Califano: We should have a charter project to support "humid" research. (seconded by Saltz)

Peterson: We need to add a participatory medicine element.

Varmus: How can we instantiate these suggestions? Think about how we get the right mix of people and mechanisms—the operational element is to be linked to people doing work.

Becich: We need a broader partnership with NLM.

Kesselman: Do not turn caGrid into caCloud.

Masys: We need a strategic plan; caBIG never had a documented plan.

Session 2: Clinical and Translational Science

Presentation 1: Dr. Abernethy, Connecting Science and Research to Clinical Care through Informatics

(NOTE: She attended and referenced the recent Institute of Medicine Cancer Informatics meeting, which discussed similar issues and ideas.) Our objective is to create a better approximation between research

and the clinic in order to facilitate clinical and translational science. In regard to creating a rapid-learning health-care system, we need a new integrating workforce culture to address the disconnect between clinicians and informaticists—we need to train clinicians and researchers to be informatics savvy. A rapid-learning health-care system depends on the ability to continuously aggregate data in an integrated way. Critical tasks to achieve a rapid-learning system include data linkage across numerous dimensions; the ability to plan, test, and learn with analytics built into the system; and the need to accommodate real-time data use and reuse. Person-centricity—most specifically, the concept of the Cancer Information Donor—will be central to the new system, as will be governance and an integrated informatics infrastructure across clinical research domains. Interoperability will be critical for integration across domains. We must keep in mind the longitudinality of clinical data—patients must be assessed over time, not in a series of disconnected snapshots. Clinical research data must be harmonized. We need to build a culture of trust, with data perceived as a public good. We must consider issues of data replicability, provenance, and sense-making. We need to catalyze with NLM and other government agencies.

Presentation 2: Dr. Dalton, A Partnership to Develop a National Health and Research Information Exchange (NHRIE)

Pathologists are central to ensuring data quality and appropriate clinical use. Data collection processes need to be examined, and we need to develop processes to manage and use "dirty" data. More data need to be captured in EMRs. There is a profound need for natural-language processing because EMRs are not based on standards; they are a mix of structured and unstructured data. To develop standards, we need to codify SOPs and utilized a shared-governance system to encourage community involvement. From the Cancer Centers' perspective, there is a need to develop a federated model, a hub-and-spoke structure, for national data exchange. The Moffitt Total Cancer Care Protocol, which has established a data warehouse that contains data on 87,000 patients and uses the Moffitt informatics platform and Oracle, provides an example of a new federated model for research and health care. The Total Cancer Care protocol centers on three questions: "can we follow you throughout your lifetime; can we study your tumor using molecular technology; can we recontact you?" Total Cancer Care has created four data portals (research view, patient view, administrators view, and clinician view) to expedite data access. The infrastructure includes an integrated data warehouse; but unlike caBIG, it does not employ a grid-based infrastructure. Public-private partnerships are essential to support this kind of project.

Presentation 3: Dr. Levy, The Fourth Decade of Cancer informatics

We are in the fourth decade of cancer informatics and in the midst of an explosion of novel biomarkers. Gaps include the need to create novel clinical decision-support applications that can handle such biomarkers and the need for better data-mining capabilities. There is an inherent tension between informatics as a research program and informatics as a service. The collaborative Quantitative Imaging Network (QIN) is an example of a successful approach that focuses on imaging biomarkers and whose system architecture is a linked data repository. Each member organization has its own research grant and its own disease, imaging modalities, and imaging biomarkers, and they collaborate to define best practices and develop shared tools. QIN's ultimate goal is to improve algorithm validation, biomarker

qualification, and clinical decision support. However, a central challenge to collaborative research and data sharing is the fact that there is no central coordinating center for standards. Perhaps NCI should push for standards surrounding data mapping and rules. Other possible roles for NCI include supporting infrastructure development, workflow repositories, research into clinical decision support and data mining, and training in cancer informatics.

Presentation 4: Dr. Saltz, Cancer Clinical and Translational Informatics Goals

CTSA is a model that illustrates interoperability needs. A major issue is informatics research to support clinical use. We need to develop methods for integrative analysis, disease subtyping, and targeted treatments based on these, as well as methods that are based on genomic, proteomic, epigenetic, tissue, and *in vivo* imaging information, with an emphasis on including quantitative analysis of radiology and pathology imaging data. We need tools for extracting inferred clinical phenotypes (clinical course, treatment, comorbidities, etc.) from EMRs. EMR systems contain rich data but are not designed to support research studies and are not readily interoperable. The NHGRI-supported Electronic Medical Records and Genomics (eMERGE) network and the Emory-supported Minority Health Genomics and Translational Research (MH Grid) network are examples of efforts to integrate omics data and clinical data to provide precise and nuanced descriptions of disease. Researchers need easier access to biomedical-friendly, high-performance computing to achieve data integration on the scale needed, as well as a powerful, flexible security architecture. Ultimately, we need to establish a learning health-care system that couples active learning with analytics.

Dr. Komatsoulis, Summing-up Remarks and Ensuing Discussion

NCI needs to step up to standards. The informatics infrastructure needs to include a public library of clinical decision-making tools. What can NCI do in this area? What about transmission standards?

Unidentified Speaker: In cancer we have two "jewels" that offer tremendous resources to researchers. This first is the vocabulary-services group at Pittsburgh, which should be preserved: their resources and services are extremely valuable. Our cancer registries (North American Association of Central Cancer Registries [NAACCR]) are also an extremely important source of data. However, the meta-thesaurus and EVS are not very well connected and need to be strengthened for use by the Cancer Centers: we need to connect them into genomics.

Becich: We need to connect phenotype to genotype in a more concerted way: an eMERGE version of cancer.

Prindiville: We need NCI systems to interoperate, and we need incentives for doctors and patients to participate.

Saltz: We cannot overstress the importance of the CTSA model. We should not forget about the *in silico* centers: a lot has been built that we can go forward with. The *in silico* centers are modeled as mini National Centers for Biomedical Computing (NCBC), a network that, among other activities, facilitates

collaborative research. We don't need reinvention; we need to look at NCBCs, *in silico* centers, and CTSA etc. as learning opportunities.

Siegel: Use the NLM model for interoperable databases.

Unidentified Speaker: We are going to have more complex trials with more correlative scientific components. These can provide more structured clinical data than any other source, but we need to create links for trans-data analyses. These can inform data exchange and be interrelated as a large, federated system.

Masys: NCI needs to reestablish novel, bold leadership to change the world in a sustainable fashion. We need to develop an infrastructure to support precision medicine and a learning health-care system, including a public library of decision-support tools.

Murphy: We should consider starting with something that works locally: a driving biological problem such as "accrual for clinical trials."

Session 3: Informatics Infrastructure: Interoperability and Data Standards

Presentation 1: Dr. Payne, The Multiple Dimensions of Interoperability and Data Standards in the Clinical and Translational Research Domains

Standards for data sharing are not uniform, and complex social issues—multiple human factors—must be taken into account. There is a need for dynamic models: evolution is desirable in complex systems. We have a need for a shared semantics. There is an important distinction between computable interoperability (*a priori*) and working interoperability (peer to peer, developed on the fly). There should be a place for both in NCIP; we have to achieve a balance between the two for agility and adaptability. We must also empower knowledge workers: real-world driving biological problems plus SMEs equal solutions to real-world interoperability needs. The OSU Translational Research and Data Management Grid (TRIAD) project exemplifies a lightweight, inexpensive model that has achieved working interoperability among a number of CTSA sites; it was developed and delivered in three months. TRIAD technology and its approach to flexible interoperability support the Hairy Cell Leukemia Research Consortium.

Presentation 2: Dr. Huff, An "App Store" for Health Care

From the Institute of Medicine: A learning health-care system is "one in which progress in science, informatics, and care culture align to generate new knowledge as an ongoing, natural by-product of the care experience, and seamlessly refine and deliver best practices for continuous improvement in health and health care." To accomplish this, there is a need for standardized APIs. This is a goal of the Strategic Health IT Advanced Research Projects (SHARP) initiative led by the Office of the National Coordinator (ONC), which aims to develop a framework of open-source services to transform EHR data into standards-conforming, comparable information suitable for large-scale analyses, inferencing, and integration of heterogeneous health data. Our needs include a standard set of detailed clinical data models, standardized APIs, standard coded terminology, and a shared repository to allow open sharing

of clinical data models, coded terms, and APIs. Creating publicly shareable standard APIs for core information that any application can use is one of his group's primary goals.

Presentation 3: Dr. Chute, Fundamental Basis for Data Standards

Standards enable dynamic translation, as shown by the eMERGE and SHARP projects. There is also a critical need for clinical data normalization and natural-language processing, both of which are being addressed by the Clinical Information Modeling Initiative (CIMI) based at the Mayo Clinic. CIMI's overarching goal is to improve interoperability of health-care information systems through shared implementable clinical information models. Interoperability and standards are crucial to achieve high-throughput phenotyping. Clinical data harmonization is crucial (CIMI). Parochial standards are disruptive. NCI must partner with other standards organizations such as Health Level 7 (HL7) and the Clinical Data Interchange Standards Consortium (CDISC). caDSR should focus on registration of these standardized elements. Additionally, NCI must collaborate with the meaningful-use community.

Presentation 4: Dr. Kush, Standards Enabling the Conduct of Clinical Research

Desired qualities of standards include being global, open, and free, as well as being developed through recognized standards-development processes and being collaborative, cooperative, and consensus based. Clinical research standards should also be harmonized and semantically consistent to support interoperability with health care. A prime example for such standards development is the Research Electronic Data Capture (REDCap) project, but its system is not easy to use. The Biomedical Research Integrated Domain Group (BRIDG) provides an example of development taking place through an iterative, moving process, as does the Clinical Data Acquisition Standards Harmonization (CDASH) initiative The Shared Health and Research Electronic Library (SHARE) pilot project, supported by CDISC, demonstrates the need for a global, accessible electronic library that enables the standard, precise definition of data elements. NCIP should invest in a new, open BRIDG/EVS-based semantic infrastructure environment. NCIP should also promote the use of relevant existing standards throughout NCI. CDISC will continue to collaborate with NCI and NCIP.

Dr. Komatsoulis, Summing-up Remarks and Ensuing Discussion

Today is the beginning of our discussions, not the end. We hope that all of you will continue to engage with us. The floor is open for a discussion of standards and interoperability.

Getz: We need translation layers between standards.

Chute: SHARP provides an approach to open, linked data.

Payne: We need a lightweight approach to start meaningful data sharing; we don't need a system that is heavyweight all the way across.

Madduri: We need to prioritize the top ten killer use cases.

Peterson: We need international collaboration around standards for creating patient portals; we do not want portals silo'ed around diseases.

Unidentified Speaker: SPOREs need biomedical informatics expertise built in.

Unidentified Speaker: Standards change; versions change. We will need help to translate data sets. We need to request that people do or put content into the ecosystem. People will not go to one standard.

Cimino: Give them a standard to work with. People want a foot in the door. NCI can begin putting out standards at the meta level.

Appendix: Citations of specific agencies/programs/initiatives and papers that could serve as exemplars or as partners as NCIP develops

- The National Centers for Biomedical Computing (NCBCs) http://www.ncbcs.org/), funded as an NIH Roadmap project, exemplify the driving biological project model.
- Clinical and Translational Science Awards program (CTSA) (the most frequently mentioned example during the meeting)—a model for the "start small" approach; CTSA also developed informatics requirements with teeth for grants https://www.ctsacentral.org/.
- The National Library of Medicine (NLM) has developed interoperable databases http://www.nlm.nih.gov/biomedical.html.
- Google—ability to handle and parse enormous amounts of unstructured data every day—lightweight service to search/analyze data but with heavyweight infrastructure.
- National Security Agency (NSA): ability to handle and parse enormous amounts of unstructured data http://www.nsa.gov/ --a lightweight service to search/analyze data but with heavyweight infrastructure.
- Encyclopedia of DNA Elements (ENCODE) project supported by NHGRI could provide a paradigm for an ENCODE+ project for cancer http://www.genome.gov/10005107.
- Cancer Genomics Hub (CGHub) https://cghub.ucsc.edu/, University of California Santa Cruz, provides a good, but limited model for data storage.
- cBio Cancer Genomics Portal (cBioportal) http://www.cbioportal.org/public-portal/ developed by Sloan Kettering exemplifies the kind of data portal needed—an open portal for exploring multidimensional cancer genomics data.
- The Cancer Genome Atlas (TCGA) data portal http://cancergenome.nih.gov exemplifies the need to integrate disparate data types.
- Integrated Genomics Viewer http://www.broadinstitute.org/igv/ exemplifies the need to integrate data.
- The electronic Medical Records and Genomics (eMERGE) network http://www.genome.gov/27540473, supported by NHGRI, demonstrates that biologists and clinicians can meaningfully interact to integrate data generated in the two domains.
- The Bionimbus Cloud project (http://www.bionimbus.org/), developed at the University of Chicago, exemplifies an open-source cloud-based system designed to support next-generation sequencing and integrates technology for analyzing and transporting large data sets.
- The World Wide Web Consortium (W3C) http://www.w3.org/, an exemplar for defining standards.

- Moffitt Total Cancer Care protocol http://www.moffitt.org/totalcancercare , an exemplar for a multidimensional data warehouse used for both research and care.

- The NCI-supported Quantitative Imaging Network (QIN) for Evaluating Responses to Cancer Therapies http://imaging.cancer.gov/programsandresources/specializedinitiatives/qin exemplifies the strengths of a collaborative research effort.

- The Electronic Medical Records and Genomics (eMERGE) network http://www.genome.gov/27540473, supported by NHGRI, provides an example of an initiative focused on integrated DMT clinical data with basic research data.

- The Minority Health Genomics and Translational Research (MH Grid) network http://cci.emory.edu/cms/projects/mh-grid.html, supported by Emory University, provides an example of an initiative focused on integrated DMT clinical data with basic research data.

- The North American Association of Central Cancer Registries (NAACCR) http://www.naaccr.org/ can prove a rich source of data to support genotype-phenotype linkage.

- For an overview of the OSU TRIAD project, see Fleming LK, Buetow KH, Payne PRO, Extending the services-oriented architecture of the cancer biomedical informatics grid—the OSU Translational Research and Data Management Grid project, *iHealth Connections*, 2012;2(1):70-74.

- The Hairy Cell Leukemia Research Consortium http://www.hairycell.org/content/consortium/welcome is supported by OSU TRIAD technology.

- The Strategic Health IT Advanced Research Projects (SHARP) initiative http://informatics.mayo.edu/sharp/index.php/Main_Page led by the ONC aims to develop a framework of open-source services to transform HER data into standards-conforming, comparable information suitable for large-scale analyses, inferencing, and integration of heterogeneous health data.

- The Clinical Information Modeling Initiative (CIMI) http://informatics.mayo.edu/CIMI/index.php/Clinical_Information_Modelling_Initiative, which has the overarching goal of improving interoperability of health-care information systems through shared implementable clinical information models.

- Health Level 7 (HL7) http://www.hl7.org/ is an authority on standards for interoperability of health information technology.

- The Clinical Data Interchange Standards Consortium (CDISC) http://www.cdisc.org/ has established standards to support the acquisition, exchange, submission, and archiving of clinical research data and metadata.

- The Research Electronic Data Capture (REDCap) http://project-redcap.org/ consortium is focused on supporting a web application designed to support data capture for research studies,
- The goal of the Biomedical Research Integrated Domain Group (BRIDG) http://www.bridgmodel.org/ is to produce a shared view of the dynamic and static semantics for the domain of protocol-driven clinical research and its regulatory artifacts.
- The Clinical Data Acquisition Standards Harmonization (CDASH) initiative http://www.cdisc.org/stuff/contentmgr/files/0/9b32bc345908ac4c31ce72b529a3d995/misc/cdash_std_1_0_2008_10_01.pdf , whose purpose is to describe basic standards for the collection of clinical trial data.
- The Shared Health and Research Electronic Library (SHARE) pilot project http://www.cdisc.org/cdisc-share, supported by CDISC, demonstrates the need for a global, accessible electronic library that enables the standard, precise definition of data elements.
- The Innovative Medicines Initiative (IMI) http://www.imi.europa.eu/ represents a potential collaborator.